Structure-aware Feature Disentanglement with Knowledge Transfer for Appearance-changing Place Recognition

Noname manuscript No.
(will be inserted by the editor)

Structure-aware Feature Disentanglement with Knowledge Transfer for Appearance-changing Place Recognition

Cao Qin · Yunzhou Zhang* · Yingda Liu · Delong Zhu · Sonya Coleman · Dermot Kerr

Received: date / Accepted: date

Abstract Long-term visual place recognition is challenging as the environment is subject to drastic appearance changes across different temporal resolutions such as time of the day, month, and season. A wide variety of existing methods address the problem by means of feature disentangling or image style transfer, but ignore the structural information which often remains stable even under environmental condition changes. To overcome this limitation, this paper presents a novel Structure-aware Feature Disentanglement Network (SFDNet) based on knowledge transfer and adversarial learning. Explicitly, probabilistic knowledge transfer (PKT) is employed to transfer knowledge obtained from the Canny edge detector to the structure encoder. An appearance teacher module is then designed to ensure the learning of appearance encoder does not only rely on metric learning. The generated content features with structural information are used to measure the similarity of images. We finally evaluate the proposed approach, and compare it to state-of-the-art place recognition methods using six datasets with extreme environmental changes. Experimental results demonstrate the effectiveness and improvements achieved using the proposed framework.

Keywords Visual place recognition, Knowledge transfer, Changing environment, Representation disentanglement

* Corresponding author.
C. Qin, Y. Zhang
College of Information Science and Engineering, Northeastern University, Shenyang, China
E-mail: zhangyunzhou@mail.neu.edu.cn(✉)

1 Introduction

Visual place recognition (VPR) involves helping robots reduce pose drifts as well as accumulated errors in trajectory by determining if the robot has been in the location previously. In the last decade, significant progress has been made in the field of vision-based simultaneous localization and mapping (SLAM) [34,16] and visual place recognition forms a key component in loop closure. However, place recognition in the presence of varying environmental conditions (such as changes in weather, illumination and season) remains extremely challenging. Existing VPR methods that use handcrafted features to generate descriptors are often hampered by the difference in features across various images sets. Once this key issue is addressed, more accurate comparison among images will be achievable. This paper concentrates on the problem of extracting invariant feature representations from image sets across varying and contrasting appearances for visual place recognition.

Benefitting from recent progress in feature representation learning based on convolutional neural networks (CNNs), CNN-based descriptors have been studied to improve the performance of place recognition tasks [8, 42,53,54]. Some works are devoted to supervised feature learning [8,9,13], but the issue holding them back currently appears to be the lack of suitably large labelled datasets. Therefore, self-supervised learning approaches [38,47,27,46,2] are deemed to be more suitable for such a task.

Recent self-supervised learning methods for VPR, which deal with extreme changes in appearance, are generally divided into two approaches. The first approach is to transfer the style of queried images (source domain) to the style of reference images (target domain) by using Generative Adversarial Networks (GANs)

[20]. These works [27,2,46] achieved realistic transformations, but they are only able to handle matching between two image domains. The second approach is to extract the invariant features from the images, regardless of the environmental condition changes. By finding latent feature spaces and separating appearance-relevant and appearance-irrelevant features, these methods [60,47,56] improve the performance of visual place recognition with respect to varying environmental conditions. Nevertheless, it is difficult to observe and understand what information the invariant features encode from the perspective of transparency and interpretability. Based on this consideration, this paper proposes an approach which guides the encoder to incorporate the information desired through knowledge transfer, thus reducing the uncertainty of invariant feature space and further improving the discriminative ability of features.

Ideally, a place recognition algorithm should have the ability to capture appearance-invariant features to cope with changes in various environments. We find that even if the appearance of scene objects (e.g., streets, houses, and trees) varies with the changes of the environment, the structure of these objects always remains stable, such as the outline and edges. Thus, we intend to use additional knowledge to assist the encoder to capture the structural information in the process of feature disentanglement. Knowledge transfer, which was presented to boost the performance of the lightweight neural network model, is usually used for transferring knowledge from a complex teacher model to a smaller and simpler student model by imitation. However, it is challenging to apply knowledge transfer to the proposed system as we intend to use an unsupervised approach to learn and separate the required features; existing knowledge transfer approaches are mostly designed for classification problems and usually require a large number of labelled target and source data. Probabilistic Knowledge Transfer (PKT), proposed in [44,45], works by matching the probabilistic distribution of the data in the feature space instead of the actual output of the network. More importantly, this method was able to transfer the knowledge from handcrafted features such as Histogram of Oriented Gradients (HoG) [11] and Local Binary Patterns (LBP) [43]. Inspired by this work, we use PKT to enable the designed encoder to learn from other teacher models. Explicitly, we use the Canny edge detector [7] to extract the edge information from the image as a teacher to guide the structure encoder to learn the feature distribution of edges.

In view of the above consideration, we propose a Structure-aware Feature Disentanglement Network (SFDNet) based on knowledge transfer and adversarial learning. Building on previous work [47] which achieved the disentanglement of appearance and content features through the use of an auto-encoder and adversarial learning, we introduce the probabilistic knowledge transfer technology to guide the structure encoder to learn more distinguishable features with structure information. Moreover, we add an additional appearance teacher module, which makes the learning of an appearance encoder less reliant on metric learning. Finally, the deep features generated by the structure encoder will be used for place recognition. Fig. 1 shows the overall framework of the proposed method and the main contributions of this work are summarized as follows:

- We develop a novel Structure-aware Feature Disentanglement Network (SFDNet) based on adversarial learning and knowledge transfer, which learns deep disentangled feature representations with structure information for place recognition.
- We achieve the transfer of structure information from the Canny edge detector to the structure encoder by introducing the PKT method. An appearance teacher model is added to help the appearance encoder generate more specific features.
- Extensive experiments are carried out to validate the effectiveness of the proposed SFDNet. The results show that the SFDNet has improved performance compared with the original FDNet, and also achieves state-of-the-art AUC-PR curves and EP values using severe condition-variant datasets.

2 Related Work

2.1 Visual Place Recognition

Visual place recognition is usually treated as an image retrieval task [10], where the problem of recognizing location is resolved by finding individual images of the most similar place. Traditionally, this has been addressed through extracting handcrafted local feature descriptors like ORB [49] and SIFT [32]. Such local features are then converted to a global descriptor using aggregation approaches such as VLAD [3] and DBoW [17]. Although these methods show excellent efficiency and deal with general situations well, few of them are found to be robust across a range of scenes with extreme appearance variation.

The general successes of deep neural networks in the field of computer vision has motivated researchers to investigate how to integrate CNNs into existing visual place recognition. In early work [55], the middle layer of a pre-trained AlexNet [25] was found to be robust against changing appearance, however its performance

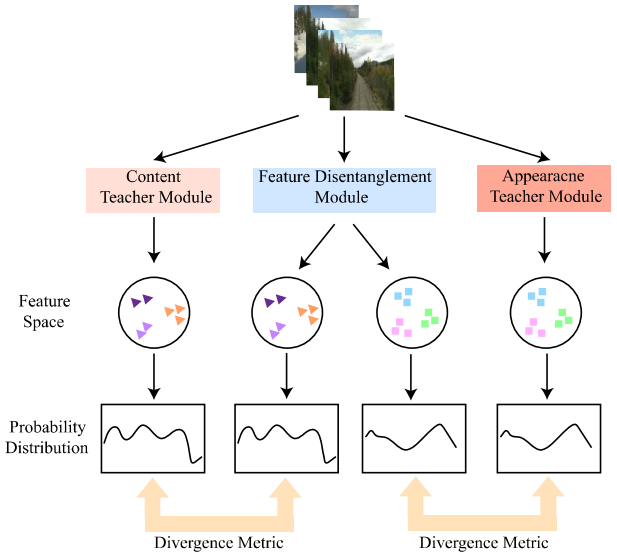


Fig. 1 Simplified principle scheme of the proposed framework. Images are decoupled into two student feature spaces: content and appearance feature spaces. The knowledge of the two teacher modules are modeled using probability distributions. Then, the knowledge is transferred by minimizing the divergence between the probability distribution of the teacher modules and the student modules.

was not satisfactory as reported in [39]. Chen et al. [8] considered appearance-invariant VPR as an image classification task. This work showed the potential for CNNs to distinguish between different locations but relied on precise and expensive human labeling. More recently, an architecture was presented that trained with the descriptors of full images in an end-to-end manner. Such a method viewed the output of CNNs as descriptor vectors that are subsequently clustered or optimized like NetVLAD, [4] and [31], resulting in state-of-the-art performance on various datasets.

To address the issue of VPR with extreme appearance changes, transforming images [31,27,2,46] from the source domain to the target domain can achieve good performance. Nevertheless, such methods fail to cope with the continuous change of appearance. Thus, extracting invariant features [60,56,47] that are robust to changing conditions is preferable. The work in [60] presented a multi-domain feature learning method, which applied a feature separation module to indirectly separate condition-related and condition-invariant features. Tang et al. [56] and Qin et al. [47] disentangled the place and appearance features using adversarial networks and self-supervised learning. However, it is not possible to discern from these frameworks what information is contained within the latent invariant features. Based on this weakness, we explore the inclusion of edge information in the image by utilising PKT, so as to further improve the robustness of the extracted invariant features explicitly.

2.2 Representation disentanglement

Disentangling the hidden factors of variations in CNN features has been widely applied for the derivation of discriminative features and for the generation of synthesized images. Liu et al. [30] proposed to disentangle the identity and attributes of a face in order to generate new face images. Fader Networks [26] learned attribute-invariant latent representations and generated variations of images using sliding attributes with the aid of an encoder–decoder architecture. To achieve image-to-image translation, Lee et al. [29] proposed a domain-invariant space, capturing shared content information across multiple domains and a domain-specific attribute space for producing diverse outputs. There are also many applications using disentangled representations to address image retrieval [37], image deblurring [35], person re-ID [12,63,18] and Audio-Visual Representations [64]. Motivated by these approaches, which have obtained promising performance in different domains, we adopt a similar approach for decoupling the image into content features and appearance features within highly dynamic scenes. Given that place recognition has no clear label for location identity, which differs from a face recognition task, in this paper we propose to generate more distinguishing features without supervision and labeled data via a structure-guided content encoder.

2.3 Knowledge Distillation/Transfer

Knowledge distillation/transfer [23] is designed for transferring knowledge from a complex teacher model to a simple student model by imitation. Generally, the knowledge is transferred by forcing the student model to regress the output of the teacher model. A large portion of proposed knowledge transfer methods [57,23,58,15] use soft-labels for the teacher model to guide the learning for the student model. However, visual place recognition is a different identification problem, where the identities of the location are infinite and thus soft-labels-based distillation approaches are not suitable. Some methods also consider using information instead of soft-labels as knowledge. Fitnets [48] trained the student model utilizing feature maps from the intermediate layers as well as soft-labels. Similarly, the Flow of Solution Procedure (FSP) [59] exploited feature maps and the flow of solution procedure (FSP) matrix for

knowledge transfer. Nonetheless, neither of these methods are applicable for the proposed task as they can not transfer knowledge from handcraft features.

Probabilistic knowledge transfer (PKT) in [44] employed a soft probabilistic distribution of data defined on the feature representations of the output layer. This work inspires us to consider using the output of the content encoder to imitate the output of the Canny edge detector. The aim is that the distribution of the features generated by the content encoder will be infinitely close to the distribution of the features extracted by the edge extractor. In this way, the structured edge information can be migrated to the content feature. Note that, in contrast to the original PKT using KL divergence as the divergence metric between distributions, we apply the Wasserstein distance measure [50] (also called EMD) which can reflect the difference between two distributions even if the support sets of two distributions do not overlap or have insignificant overlap.

3 Structure-aware Feature Disentanglement Network

We introduce the proposed Structure-aware Feature Disentanglement Network (SFDNet) based on adversarial learning and knowledge transfer. The goal is to make the content features identical across different scene appearances, and retain the structural information from the image as much as possible. We first present the overall architecture and then elaborate on the design of the feature disentanglement module, the content teacher module, and the appearance teacher module.

3.1 Overview

Finding an appropriate feature space for images is the core task in the proposed architecture. In previous work [47], we achieved feature disentanglement by using adversarial learning, where the images are decoupled into content and appearance features. In this paper, we assume that the content vector should contain structural information. As shown in Fig. 2, the feature disentanglement module consists of a structural content encoder E_{SC}, an appearance encoder E_A, an appearance discriminator D_{AA} and a decoder D_E. Given a mini-batch of images X, the structural content vector X_{SC} and the appearance vector X_A will be generated respectively. X_{SC} is fed into D_{AA} to distinguish whether the extracted structure content vectors are from the same domain. In addition, the distribution of X_{SC} will be compared with the X_{CT} which is generated by the content teacher module. This step aims to transfer the structural information to the content vector. As for X_A, not only will it be optimized by triplet loss, but it's distribution will also be compared with X_{AT}, derived from the fully connected layer of the appearance teacher module. Finally, the decoder D_E reconstructs images from the fused features to encourage learning of the content and appearance features to form a full representation of the input image.

3.2 Feature Disentanglement Module

3.2.1 Appearance Encoder Loss

As illustrated in Fig. 2, the appearance features are extracted by the encoder E_A, written as $X_A = E_A(X)$. Intuitively, the appearance encoder should capture the appearance information. Thus, for appearance features from the same domain the distance between them should be small, while for appearance encodings from different domains the distance between them should be further and greater than a given threshold. In this way, we train the appearance encoder by the triplet loss function [51]:

$$L_{AE}\left(\theta_A\right) = \sum_{\substack{a,p,n \\ y_a=y_p\neq y_n}} \left[m + D_{a,p} - D_{a,n}\right]_+ \tag{1}$$

where m is a margin that is enforced between positive and negative pairs, $D_{a,p}$ and $D_{a,n}$ are the Euclidean distances between the anchor and the positive/negative samples respectively, y_a, y_p and y_p are the appearance ID of the anchor, positive and negative samples, respectively and θ_A is the parameter of the appearance encoder.

3.2.2 Structural Content Encoder Loss

To obtain appearance-independent features, we design an approach to discriminative appearance classification loss using adversarial supervision. At the training stage, we send content encodings into the appearance discriminator D_{AA}. The purpose of the encoder E_{SC} is to fool D_{AA} so that it can not classify content features correctly.

On one hand, we need to train the appearance discriminator D_{AA} based on the generated features X_{SC} and supervise using the cross-entropy loss:

$$\begin{aligned} L_D^{adv}\left(\theta_{DAA}\right) = -\sum_{(x_i,x_j)\in X_{SC}} & y * \log\left(D_{AA}\left(x_i, x_j\right)\right) \\ & + (1-y) * \log\left(1 - D_{AA}(x_i, x_j)\right) \end{aligned} \tag{2}$$

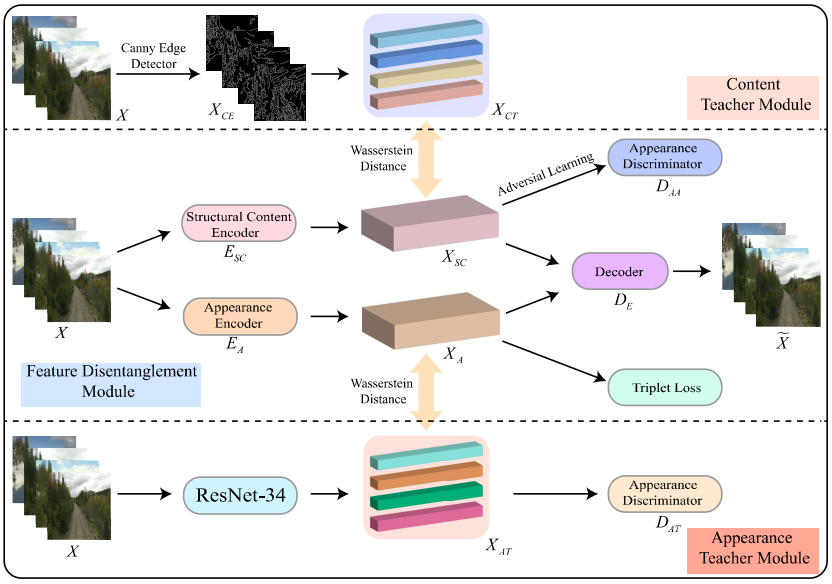


Fig. 2 The pipeline of the proposed method. In summary, there exist three core modules: 1) a Feature Disentanglement Module that is responsible for separating the structure content features and appearance features; 2) a Content Teacher Module to extract edge information and transfer to the structural content encoder; 3) an Appearance Teacher Module for helping the learning of appearance encoder.

where D_{AA} is treated as a binary classifier. For each feature pair $\{x_i, x_j\}$ with it's label y, $y = 1$ indicates that x_i and x_j come from the same domain and when $y = 0$ they are from different domains. θ_{DAA} is the parameter of the appearance discriminator and $L_D^{adv}(\theta_{DAA})$ can also be simplified to:

$$L_D^{adv}(\theta_{DAA}) = \mathbb{E}_x\left[\log D_{AA}(x)\right], \tag{3}$$

where x is the concatenation of the input content feature pair $\{x_i, x_j\}$. Note that the gradients of $L_D^{adv}(\theta_{DAA})$ are only back-propagated to the classifier but do not update the preceding layers in E_{SC}, similar to the discriminator in a GAN [20]. On the other hand, we need to train the structure content encoder to trick the appearance discriminator, where the synthesized "ground truth" appearance distribution is constant across all content features, therefore the content features are fooled to make non-informative outputs over the varying appearances. Thus it is equivalent to maximizing the entropy:

$$L_E^{adv}(\theta_{SC}) = -L_D^{adv}(\theta_{DAA}) = -\mathbb{E}_x\left[\log D_{AA}(x)\right] \tag{4}$$

where the gradients for L_E^{adv} are back-propagated to E_{SC} with the weight of the appearance classifier fixed.

To stabilize the training process which may be affected by the volatile gradient, we substitute equation 3 with a Wasserstein GAN objective which contains a gradient penalty [21]:

$$\begin{aligned} L_D^{adv}(\theta_{DAA}) =& \mathbb{E}_x\left[\log D_{AA}(x)\right] \\ &+ \lambda_{gp}\mathbb{E}_{\hat{x}}\left[\left(\left\|\nabla_{\hat{x}} D_{AA}(\hat{x})\right\|_2 - 1\right)^2\right] \end{aligned} \tag{5}$$

where $\hat{x}$ is sampled uniformly along the straight lines connecting pairs of features $\{x_i, x_j\}$ which have different appearance labels and λ_{gp} is a weighting parameter.

3.2.3 Encoder-Decoder Loss

While loss functions imposed on the appearance encoder and structural content encoder encourage the decoupling of the input image representation, there is no guarantee that the integration of X_{SC} and X_A can compose a complete encoding of the inputs. As a result, an encoder-decoder architecture is applied and the reconstruction loss is defined as:

$$L_r(\theta_{SC}, \theta_A, \theta_{DE}) = \left\|X - \tilde{X}\right\|_2^2, \tag{6}$$

where $\tilde{X} = D_E(X_{SC}, X_A)$ and $\theta_{SC}, \theta_A, \theta_{DE}$ are the parameters of the encoders and the decoder respectively.

3.3 Content Teacher Module

Due to the difficulty for the encoder to disentangle image features associated with structure information, we adopt a Content Teacher model to guide the learning of the content encoder. We use the Canny Edge Detector to extract the edge representations X_{CE} of an image batch and convert them to the vectors X_{CT}.

With reference to PKT [44], probabilistic modeling of data sample sets in two feature spaces is required. In this way, the problem of how to transfer knowledge

(edge information) from X_{CT} to X_{SC} is reduced to minimizing the divergence between the joint identity probability estimates for the teacher model P and the student model Q. Considering that the conditional probability distribution represents the probability that each sample will select its neighbors [28], we are able to model the geometry of the feature space more accurately. Thus the conditional probability distribution is used to describe the Content Teacher model:

$$p_{i|j}^{(ct)} = \frac{K\left(\mathbf{x}_i, \mathbf{x}_j; 2\sigma_t^2\right)}{\sum_{i=1, i\neq j}^{N} K\left(\mathbf{x}_i, \mathbf{x}_j; 2\sigma_t^2\right)} \in [0,1] \tag{7}$$

Similarly, the probabilistic representation of X_{SC} (the student model) is denoted as:

$$p_{i|j}^{(sc)} = \frac{K\left(\mathbf{y}_i, \mathbf{y}_j; 2\sigma_t^2\right)}{\sum_{i=1, i\neq j}^{N} K\left(\mathbf{y}_i, \mathbf{y}_j; 2\sigma_s^2\right)} \in [0,1] \tag{8}$$

where $K\left(\mathbf{a}, \mathbf{b}; \sigma_t^2\right)$ is a symmetric kernel function with width σ_t, and $\mathbf{a}$ and $\mathbf{b}$ are the input vectors. The conditional probabilities sum to 1 and are bound to $[0, 1]$.

With respect to the choice of kernel function, there are several options such as Gaussian kernels, Cosine kernels and T-student kernels. Since the cosine kernel function does not require domain-dependent tuning and is widely used in retrieval tasks, a cosine similarity based metric is employed in this teacher-student model. The applied metric is denoted as:

$$K_{\text{cosine}}(\mathbf{a}, \mathbf{b}) = \frac{1}{2}\left(\frac{\mathbf{a}^T\mathbf{b}}{\|\mathbf{a}\|_2\|\mathbf{b}\|_2} + 1\right) \in [0,1] \tag{9}$$

There are also several different choices for defining the divergence metric to train the student model. In the original PKT, the Kullback–Leibler (KL) divergence is used. However, the distribution of the features generated by the encoder and the distribution of the edge features do not overlap or the overlapping part can be ignored, which will lead to the KL divergence being meaningless. Therefore, we use the Wasserstein distance as the divergence metric:

$$W\left(P_1, P_2\right) = \inf_{\gamma \sim \Pi\left(P_1, P_2\right)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|], \tag{10}$$

where P_1 and P_2 are the probability distributions of the teacher and student models respectively, $\Pi\left(P_1, P_2\right)$ is the set of all possible joint probability distributions between P_1 and P_2. As a distance function, Wasserstein distance has a valuable characteristic in that it is bounded from below by the distance between the centroids of two distributions. Adopting such a lower bound significantly reduces the number of computations required. The final loss function used for training the student model (structural content encoder) is defined as:

$$L_{CT}\left(\theta_{SC}\right) = \frac{1}{N(N-1)^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \sum_{k=1, k\neq i}^{N} \left| p_{j|i}^{(ct)} - p_{k|i}^{(sc)} \right| \tag{11}$$

where N is the batch size.

3.4 Appearance Teacher Module

To further improve the disentanglement, we additionally use an Appearance Teacher Module to guide the learning of the output features of the appearance encoder. The Appearance Teacher Module is treated as a multi-class classification model and is pre-trained on an image set with appearance labels. The penultimate layer of the neural network denoted as X_{AT} is used to extract the features for knowledge transfer. Similar to the Content Teacher Module, the Wasserstein distance is utilized to measure the similarity of the probability distribution between X_{AT} and X_A. The loss function of the appearance teacher-student model is defined as:

$$L_{AT}\left(\theta_A\right) = \frac{1}{N(N-1)^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \sum_{k=1, k\neq i}^{N} \left| p_{j|i}^{(at)} - p_{k|i}^{(a)} \right| \tag{12}$$

Algorithm 1 outlines the main steps of the calculations of the transfer loss functions L_{CT} and L_{AT}.

Algorithm 1 Calculation of a transfer loss function

Input: batch_size N, feature vectors from teacher model $\mathbf{V} \in \mathbb{R}^{N\times M_1}$, feature vectors from student model $\mathbf{W} \in \mathbb{R}^{N\times M_2}$
Output: loss: L

for $V_i \in \mathbf{V}$, $W_i \in \mathbf{W}$ **do**
  $V_i \leftarrow normalize(V_i)$
  $W_i \leftarrow normalize(W_i)$
end for
$S_v \leftarrow$ Calculate *cosine_similarity* by $\frac{1}{2}(\mathbf{V}\mathbf{V}^T + 1)$
$S_w \leftarrow$ Calculate *cosine_similarity* by $\frac{1}{2}(\mathbf{W}\mathbf{W}^T + 1)$
$P_v \leftarrow$ Calculate *probabilities* by $\frac{S_v^{i,j}}{\sum_{i=1, i\neq j}^{N} S_v^{i,j}}$
$P_v \leftarrow$ Calculate *probabilities* by $\frac{S_w^{i,j}}{\sum_{i=1, i\neq j}^{N} S_w^{i,j}}$
$L \leftarrow calcEMD(P_v, P_w)$ by equation 11 or 12
return L

3.5 Training and Testing

As shown in Algorithm 2, the training process contains two main stages. One is the encoder-decoder training stage and the other is the appearance discriminator training stage. We sequentially update the encoder, decoder, and discriminator with the following gradients:

$$\theta_{SC},\theta_A,\theta_{DE} \xleftarrow{+} -\Delta_{\theta_{SC},\theta_A,\theta_{DE}} \left(\lambda_f(L_r + L_{AE} + L_E^{adv}) + \lambda_{ct}L_{CT} + \lambda_{at}L_{AT}\right) \tag{13}$$

$$\theta_{DAA} \xleftarrow{+} -\Delta_{\theta_{DAA}} \left(L_D^{adv}\right) \tag{14}$$

where λ_f, λ_t and λ_{at} are adjustable weights. During testing, we only need to extract the structure content features for evaluation. The evaluation is performed using single-image matching based on the cosine distance of the extracted feature vectors. Specifically, given a sequence of previously visited locations, we extract their structure content features $\{F_{DB,i}\}$ through E_{SC}, where $i \in [1, N_{DB}]$ and N_{DB} is number of images in the dataset. When a new query image is available, the query features F_Q are compared with the features in $\{F_{DB,i}\}$. The best match x_t is determined by:

$$t = \arg\min_i \left(\frac{F_Q \cdot F_{DB,i}}{\|F_Q\| \, \|F_{DB,i}\|}\right) \tag{15}$$

Algorithm 2 Learning of SFDNet

Input: batch_size N , domain_num N_d , a set of training images X_s, Adjustable weights λ_f, λ_{ct}, λ_{at}
Output: parameters: $\theta_{SC},\theta_A,\theta_{DE},\theta_{DAA}$

```
θ_SC, θ_A, θ_DE, θ_DAA ← initialize;
for Iters. of whole model do
    X ← Sample mini-batch from X_s in shape (N, N_d)
    T ← Generate triplets
    P ← Generate pairs with labels by sampling from X
    for Iters. of updating encoder-decoder do
        θ_SC, θ_A, θ_DE ←+ −Δ_{θ_SC,θ_A,θ_DE} (λ_f(L_r + L_AE
                          + L_E^adv) + λ_ct L_CT + λ_at L_AT)
    end for
    for Iters. of updating discriminator do
        θ_DAA ←+ −Δ_{θ_DAA} (L_D^adv)
    end for
end for
return θ_SC, θ_A, θ_DE, θ_DAA
```

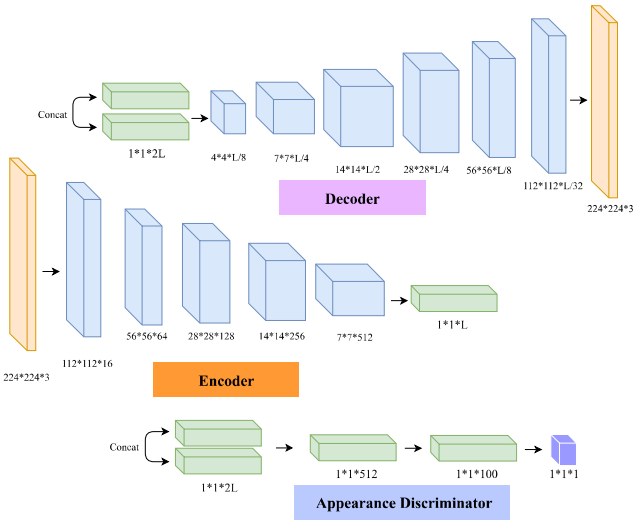


Fig. 3 Network architectures of the encoder, decoder and the discriminator

4 Implementation

Model. Firstly, for the feature disentanglement module shown in Fig. 3, we adopt the same network architecture as in [47]. The images are all resized to a uniform size (224×224) before entering the network, and the corresponding image pairs are constructed simultaneously. The content encoder and appearance encoder have the same structure, and each encoder will generate feature encodings of length L through a fully-connected layer. In order to ensure that the distribution of the structure content features and appearance features are different, the parameters of the two encoders are not shared. The decoder consists of several up-sampled blocks that contain a deconvolution layer, batch normalization and a Leaky Rectified Linear Unit. The appearance discriminator is composed of three fully-connected layers. As for the appearance teacher model, we extract the output from the penultimate layer of the ResNet-34 network [22] as the transferred features. The appearance teacher network needs to be retrained to classify the appearances of the images. Therefore, a revised classification layer with the Softmax activation function is added. Finally, for the content teacher model, since we use the Canny Edge Detector to derive the edge information, we do not need to consider the design of the network structure.

Optimization. We implement the proposed method based on Tensorflow [1]. The network is trained using a pre-trained model of FDNet [47] with the Adam optimizer, a learning rate of 1×10^{-4} and momentum of 0.5. The feature length L is set to 512. The batch size is set to 4. The margin m in the triplet loss function is set to 0.1 and λ_{gp} is set to 10. The dropout rate is

0.5 in the encoder, and 0.25 in the discriminator. The appearance teacher module is fine-tuned based on the ResNet-34 network using the Adam optimizer with an initial learning rate of 1×10^{-4} and a decay factor of 0.1 within 2 epochs. The programs run on a PC equipped with an Intel Xeon CPU at 2.10 GHz and NVIDIA TITAN X GPU with 12GB GPU memory.

5 Experiments

5.1 Datasets

To evaluate the capabilities of the proposed approach, we conduct extensive experiments using several datasets which is necessary to ensure significant appearance changes but not too much view-point change. Also, the ground truth of the corresponding locations is provided. The characteristics of each dataset are summarized in Table 1, which presents information about the type of appearance variations, the type of viewpoint variation, the tolerance of the ground truth, the number of training and testing images used and a brief description about each dataset. Sample images of these datasets are presented in Fig. 4.

5.2 Comparison Methods

To evaluate the performance of SFDNet, we compare it with several different state-of-the-art methods as outlined below:

(a) **Conv3:** Sünderhauf et al. [9] suggested that the Conv3 layer of the AlexNet is the most robust to conditional changes. The version in Python along and Gaussian random projection is used to encode the original feature maps into vectors.
(b) **NetVLAD:** NetVLAD [2] added a trainable VLAD layer to the last convolutional layer and achieved weakly VPR-specific training. We use the Pytorch implementation of NetVLAD with the online hardest triplet Loss.
(c) **CALC:** CALC [38] reconstructed similar HoG-descriptors in an unsupervised way through a convolutional auto-encoder network. We fine-tune the model using the provided open-source implementation.
(d) **HybridNet:** Chen et al. [8] re-trained the model based on the initialized weights of CaffeNet [25] specifically for place recognition. We have extracted the Conv4 layer from the HybridNet model to use as comparative features.
(e) **CoHOG:** CoHOG [61] combined Region of Interest (ROI) extraction and HoG-descriptors for regional convolutional matching for VPR. We use the open-source version and adopt a cosine distance for feature descriptor comparison.
(f) **Region-VLAD:** Region-VLAD was introduced by Khaliq et al. [24] combining lightweight CNN architectures and VLAD for VPR. We implemented it in accordance with [62].

5.3 Evaluation Metrics

To evaluate the performance of SFDNet, we use the following metrics: **PR curves and AUC:** We test the VPR performance of the proposed models using precision-recall (PR) curves and area under the curve (AUC) metrics. Generally, precision (P) and recall (R) are computed from the similarity matrix where a varied threshold is used to obtain different TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative) values. Therefore, the Precision (P) and Recall (R) values can be computed as follows:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN} \tag{16}$$

The Precision values are plotted against the Recall, and the area under the PR curve is computed, which is termed as AUC and has a value between 0 and 1 where 1 indicates perfect accuracy.
EP Value: Considering that AUC usually ignores the Recall at 100% Precision (R_{P100}) and Precision at 0% Recall (P_{R0}), which are essential for some practical applications [33,6], Ferrarini et al. [14] presented a generic evaluation metric, namely 'Extended Precision (EP)', that takes into account R_{P100} and P_{R0} simultaneously. EP value is calculated as:

$$EP = \frac{P_{R0} + R_{P100}}{2} \tag{17}$$

5.4 Experiments using Different Parameters

In order to investigate the impact of the parameters λ_f, λ_{ct} and λ_{at} in Equation 13, we test the performance on the spring-winter sequences of the Nordland dataset. We first set λ_f to 1 and then adjust the ratio of λ_{ct} and λ_{at}. Table 2 and Fig. 5 present the results. It can be observed that when $\lambda_{ct} = 5$ and $\lambda_{at} = 1$, the proposed model achieves the best EP value and the second highest AUC.

5.5 Experiments using Different Sensitivity Thresholds

We use the edge detection algorithm as a teacher model to guide the learning of the content encoder, therefore

Table 1 Main characteristics of the datasets used in the experiments.

Dataset	Appearance Variation	Viewpoint Variation	Environment	Ground Truth Tolerance	Trained Images	Tested Images	Description
Nordland [52]	Seasonal	Small	Train Journey	5 frames	27000	1000	A dataset built from footage for each season along a railway. Ground truths are available from GPS records.
Alderley [40]	Day-Night	Moderate	City-like	2 frames	13607	1000	Two sequences recorded in the same routes in Alderley across a sunny night and a rainy day.
Gardens Point Walking [8]	Day-Night	Moderate	University Campus	3 frames	140	60	Three traverses along the same route, one during the night (right side) and two during the day (left and right side). The frame correspondences have been provided.
St Lucia [19]	Morning-Afternoon	Moderate	City-like	30 meters	3500	500	A dataset captured by a car moving through a suburb in Brisbane at different times of a day. GPS readings are provided.
Oxford RobotCar [36]	Dawn-Dusk	Moderate	City-like	30 meters	1758	754	100 repetitions of car traverse through Oxford, recorded over a year at different times of day.
FAS [41]	Summer-Winter	Moderate	City-like	3 frames	3130	1347	Three sequences recorded by a camera-equipped car in Freiburg across different seasons including summer and winter.

Fig. 4 Sample images for all evaluated datasets.

Table 2 Results on different parameters λ_{ct} and λ_{at}.

λ_{ct}	1	1	1	1	1	1	0.1	0.5	1	2	**5**	10
λ_{at}	0.1	0.5	1	2	5	10	1	1	1	1	**1**	1
EP	0.7155	0.7215	0.728	0.72	0.714	0.713	0.704	0.711	0.728	0.733	**0.7415**	0.717
AUC	0.700	0.711	0.712	0.694	0.671	0.655	0.686	0.701	0.711	0.717	**0.714**	0.697

the parameters of the Canny edge detector are also important to the performance of the proposed model. Different sensitivity thresholds lead to different noise and accuracy of the generated structure information. We adjust the threshold t from 0.02 to 0.12 to test the effect of adding the Content Teacher module on the basis of the FDNet. The PR curve is displayed in Fig. 6(a), and the respective EP and AUC values are shown in Fig. 6(b). We find that a small threshold ($t = 0.02$) results in increased noise, thus reducing the overall performance. However large thresholds, such as 0.10 and 0.12, results in less edge information being retained, which also reduces the performance. When the threshold is in the appropriate range, such as t = 0.06, the best results are achieved.

5.6 Experiments using Different Components

Our methods consist of three components. The basic module FDNet is responsible for disentangling the content and appearance features. The content teacher module (CTM) is responsible for transferring structure information to the content encoder. The appearance teacher module (ATM) is used for guiding the learning of the appearance encoder. Removing the CTM will result in less constraint on the content features, and thus degrade the overall performance of VPR. Removing the

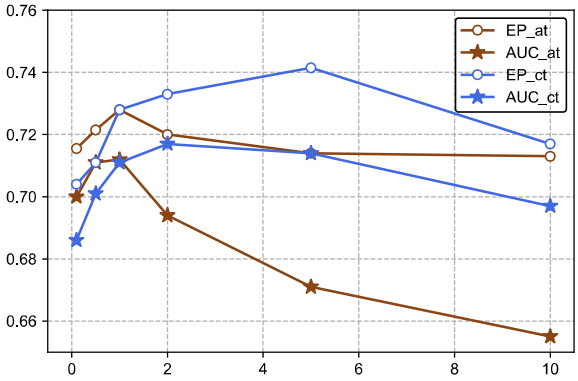
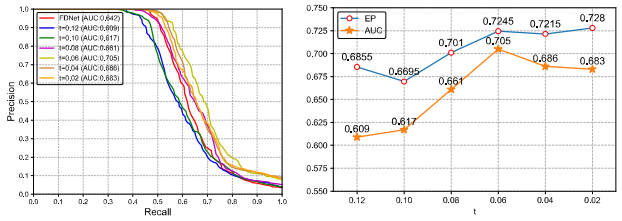


Fig. 5 Results on different parameters λ_{ct} and λ_{at}. The blue line is case when λ_{at} is fixed to 1, and the brown line is case when λ_{ct} is fixed to 1.



(a) (b)

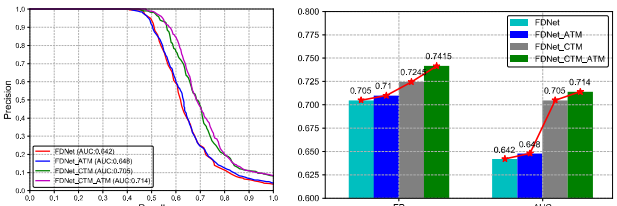
Fig. 6 Results of different sensitivity thresholds of the Canny edge detector. (a) PR curves. (b) EP and AUC values.

ATM will lead to a decrease in the ability to distinguish the appearance features, and also has a indirect impact on the recognition performance of content features.

To validate the effectiveness of the key components in the proposed framework, we conduct comparisons between different scenarios using the spring-winter sequences in the Nordland dataset. Fig. 7 shows the PR curves and the corresponding EP and AUC values. By removing both the CTM and ATM, the performance drops to 0.705 EP and 0.642 AUC as it can only rely on the original separated content features. When we remove the ATM, the performance does not drop so significantly. However, when we remove the CTM, the performance drops dramatically especially the AUC value whether or not the ATM is included. So the content teacher module is essential for the whole architecture, as it is able to boost the discriminating ability of the content features.

5.7 Experiments on Appearance Prediction

As we know, the disentangled appearance features can be used to predict the appearance characteristics of each image. We test its prediction precision on the Nordland dataset which contains 4 different visual appear-



(a) (b)

Fig. 7 Results of different components. (a) PR curves. (b) EP and AUC values.

ances. As shown in Fig. 8, the appearance features generated by the original FDNet can only achieve 70.04% mAP before the ATM is adopted. The individual ATM can achieve 91.29% mAP after fine-tuning, thanks to the deeper network structure and pre-training parameters of ResNet-34. From Fig. 8, we can also find that the addition of the CTM has little impact on the mAP for appearance features, while the addition of the ATM can improve the performance by at least 10% (70.14%→83.41%, 71.75%→ 82.33%). This means that the architecture is capable of efficiently transferring knowledge from the ATM to the appearance encoder.

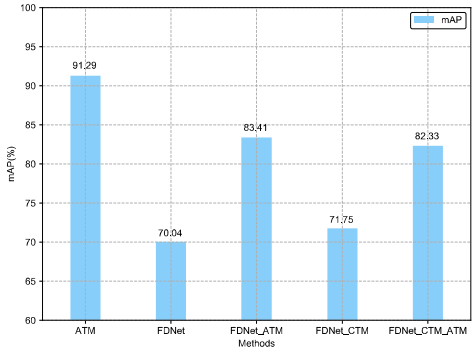
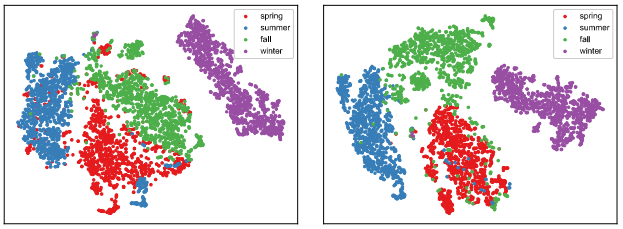


Fig. 8 Comparison with different components on appearance prediction.

5.8 Visualization of Appearance Features

We visualize the distribution of the appearance features. As shown in Fig. 9, compared with the feature distribution of FDNet, we find that the distance between the feature points of winter and the other three types of feature points is still very large, and the gaps are still obvious. The distances between feature points of summer and spring have increased, and the overlapped points become less. Moreover, the difference between the features of autumn and those of spring has been significantly improved, although there are still some

indistinguishable feature points. These phenomena show that our framework can improve the discriminative ability of appearance features.



(a) (b)

Fig. 9 Visualization of appearance features on the Nordland dataset using t-SNE. (a) FDNet. (b) SFDNet.

5.9 Experiments on Content Features

As there are hundreds of locations of content features in a dataset, we are unable to show their distribution explicitly. Therefore, the intra-location distances of content features in the same locations and inter-location distances in different locations are calculated as shown in Fig. 10. We randomly selected 100 locations in the Nordland dataset and extracted the corresponding content features. It is observed that the proposed SFDNet (FDNet_CTM_ATM) is effective in decreasing the intra-location distance, which means that SFDNet is more robust to the extreme changes in appearance than FDNet due to the addition of structure information.

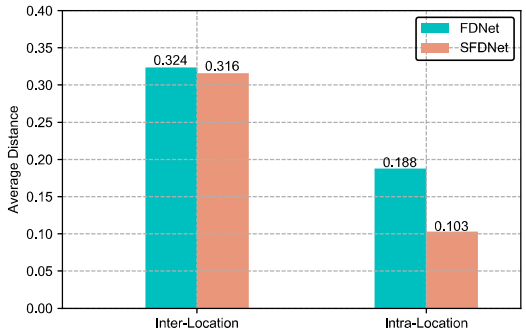


Fig. 10 The average values of intra-location distances and inter-location distance on the Nordland dataset. The experiments are respectively conducted by FDNet and SFDNet.

Table 3 Effect of the different divergence metric on ATM and CTM.

	ATM(KL)		ATM(EMD)	
	EP	AUC	EP	AUC
CTM(KL)	0.711	0.680	0.716	0.689
CTM(EMD)	0.738	0.703	**0.7415**	**0.714**

5.10 Experiments using the Different Divergence Metric

We evaluated the effect of the divergence metric used for estimating the difference of probability distributions and the effectiveness of the architecture. The results are reported in Table 3. The test sequence is also the spring-winter sequence in the Nordland dataset. We find that when ATM and CTM use KL divergence as the divergence metric, the results are poor. This may be due to the fact that the features extracted by the Canny edge detector and the features extracted by structure content encoder do not overlap in the initial distribution, leading to KL divergence losing meaning and reducing the final performance. When EMD is applied to ATM, the overall performance is not greatly improved compared with KL divergence. However, when EMD is used with the CTM, the performance is clearly improved. The best results (EP:0.7145, AUC:0.714) are obtained when EMD is used as the divergence metric with ATM and CTM.

5.11 Experiments on Different Viewpoints

Although the proposed method mainly deals with appearance changes, we still try to observe whether the proposed method has an improvement in viewpoint changes compared with the FDNet. We conduct experiments using the Nordland dataset and simulate viewpoint changes by using translated and rotated images with reference to [5]. We use 1000 pairs of images from the spring and winter seasons. Viewpoint changes are simulated by translating or rotating the images, and then cropping them and resizing to the original size. Consequently, the performances of various viewpoints between images using a translation of a 10%, a rotation of a 10% and a combination of both are compared. Fig. 11 shows the simulated viewpoints and Fig. 12 illustrates the results. We found that the changes in viewpoints negatively affect both FDNet and SFDNet. When there is only translation change, the negative effect on SFDNet is acceptable. However, when the image is rotated, the performance of SFDNet is unexpectedly worse than that of FDNet. We believe that the integration of edge information may cause the generated content feature to be

more sensitive to the rotation of the viewpoints. This shortcoming will be the focus of future research.

5.12 Comparison with State-of-The-Arts

In this subsection, we make comparisons with the state-of-the-art methods using a range of datasets. The EP and AUC are listed in Table 4, and the PR curves are presented in Fig. 13.

Nordland Dataset. Fig. 13(a) reveals that the performance of SFDNet is further improved compared with FDNet using this dataset, but there is still a gap between SFDNet and NetVLAD in terms of AUC. Region-VLAD was slightly inferior to NetVLAD, while CoHOG achieved the worst performance, even lower than Conv3. CALC shows moderate performance and the HybridNet greatly improves compared with the original Conv3 features owing to the CNN training on the VPR dataset.

Alderley Dataset. The PR curves plotted in Fig. 13(b) show a surprising result for the proposed method in this challenging case. We achieved the best performance with respect to both the EP (0.575) and AUC (0.620) values. Moreover, the downward trend of the curve is also more gradual than that of NetVLAD. CoHOG and Conv3 are still the two worst performing methods, while the others perform adequately. The extreme condition changes in this dataset, which are difficult for humans to distinguish, make it difficult for most methods to maintain stable performance. However, benefitting from the integration of structured information, our proposed approach has achieved better results.

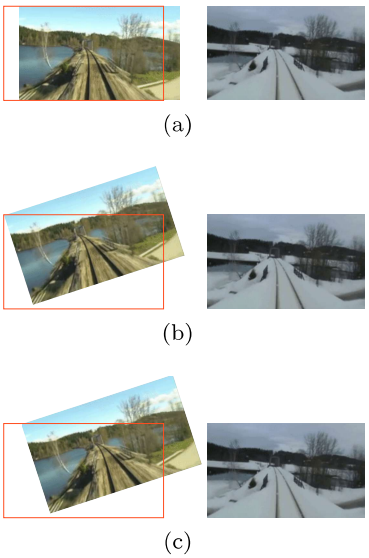


Fig. 11 Examples for the simulated viewpoint variation. (a) Translation. (b) Rotation. (c) Translation and rotation.

Gardens Point Walking Dataset. Although this dataset exhibits strong illumination, the majority of the tested VPR approaches perform relatively well as illustrated in Fig. 13(c). This is because the distinctive objects or structures are contained in both the traverses. The best results are achieved by NetVLAD (EP:0.661, AUC:0.785). The performance of Region-VLAD and CoHOG are comparable to that of NetVLAD, thanks to the information-rich objects contained in the image. Our proposed method yields adequate results both in the EP (0.628) and AUC value (0.696), this may be due to the rotation changes of viewpoints in the image sequence.

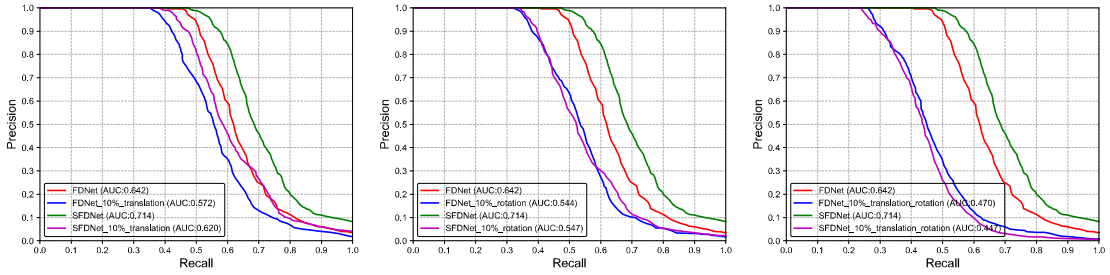
St Lucia Dataset. Using this dataset, the EP value of SFDNet is better than FDNet as expected, but the AUC value is slightly worse than NetVLAD. Condition changes are not much stronger in this dataset, therefore HybridNet, Region-VLAD and CALC perform comparably, and even CoHOG is not far behind. However, the performance of Conv3 is obviously worse than other methods, which indicates that it is not very suitable for place recognition tasks which involve extracting features directly from a model pre-trained on the object recognition dataset.

Oxford RobotCar Dataset. The PR curves on the Oxford RobotCar dataset are shown in Fig. 13(e). It is clear that NetVLAD and Region-VLAD can achieve much better results with respect to AUC values than other methods, while CoHOG and CALC follow-up with relatively poor performance. The proposed SFDNet has only a slightly higher EP value (0.570). Dynamic objects such as pedestrians and cars in this dataset become the biggest obstacle to our approach, but on the contrary, NetVLAD-based methods are good at dealing with the presence of multiple objects.

FAS Dataset. Similar to the results of the Gardens Point Walking dataset, SFDNet does not show any advantages using this dataset (only 0.609 EP and 0.694 AUC). However, our method has still achieved some improvement compared with FDNet, as far as the shape of curve and AUC are concerned. NetVLAD obtained the highest AUC value (0.760) and FDNet obtained the best EP value (0.624). CALC and Region-VLAD exhibit similar performances but approaches including Conv3, CoHOG and HybridNet have relatively underperformed.

5.13 Computational performance

In VPR systems, the computational cost is an important factor that needs to be considered when comparing

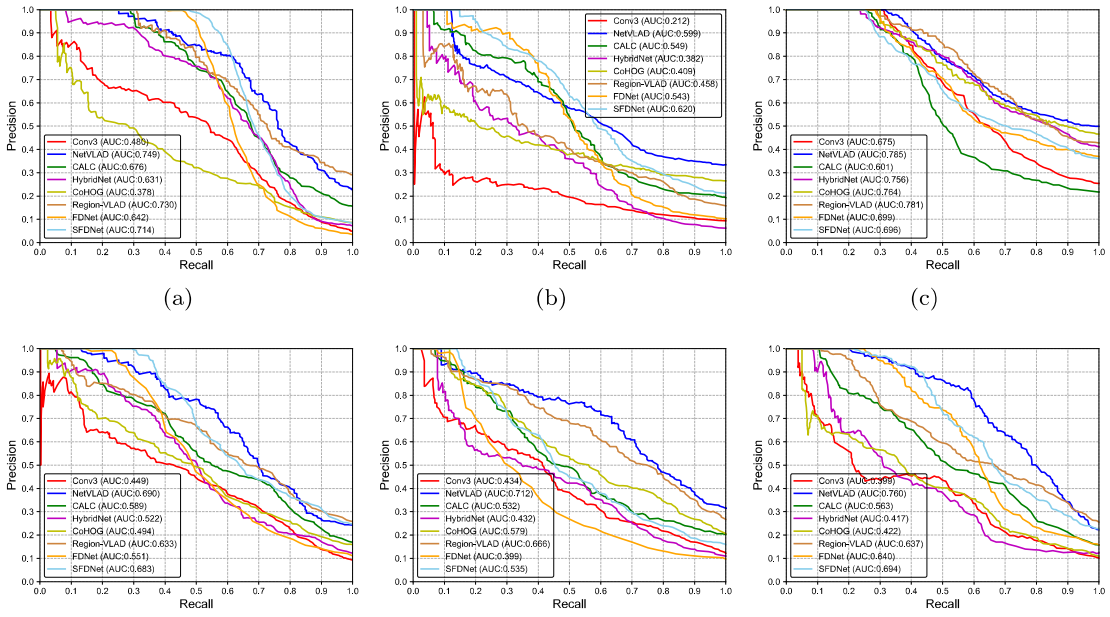


(a) (b) (c)

Fig. 12 PR curves for different viewpoint changes. (a) Translation of 10%. (b) Rotation of 10%. (c) Translation of 10% and rotation of 10%.

Table 4 EP and AUC values respectively comparing all methods on different datasets.

Method	Nordland spring-winter		Alderley day-night		Gardens Point day-night		St Lucia morning-afternoon		Oxford RobotCar night-day		FAS summer-winter	
	EP	AUC	EP	AUC	EP	AUC	EP	AUC	EP	AUC	EP	AUC
Conv3	0.517	0.480	0.502	0.212	0.656	0.675	0.501	0.449	0.513	0.434	0.519	0.399
NetVLAD	0.626	**0.749**	0.560	0.599	**0.661**	**0.785**	0.567	**0.690**	0.545	**0.712**	0.601	**0.760**
CALC	0.645	0.676	0.533	0.549	0.631	0.601	0.524	0.589	0.535	0.532	0.552	0.563
HybridNet	0.540	0.631	0.522	0.382	0.619	0.756	0.527	0.522	0.539	0.432	0.541	0.417
CoHOG	0.525	0.378	0.505	0.409	0.650	0.764	0.512	0.494	0.558	0.579	0.525	0.422
Region-VLAD	0.654	0.730	0.516	0.458	0.646	0.781	0.535	0.633	0.529	0.666	0.556	0.637
FDNet	0.705	0.642	0.554	0.543	0.644	0.699	0.572	0.551	0.536	0.399	**0.624**	0.640
SFDNet	**0.742**	0.714	**0.575**	**0.620**	0.628	0.696	**0.649**	0.683	**0.570**	0.535	0.609	0.694



(a) (b) (c)

(d) (e) (f)

Fig. 13 PR curves comparing all other approaches with our novel method on different datasets. (a) Nordland. (b) Alderley. (c) Gardens Point Walking. (d) St Lucia. (e) Oxford RobotCar. (f) FAS.

a query image against a large number of database images. Table 5 presents the time taken (in ms) for feature encoding for each of the methods and the time taken for feature matching given a single query image. We use 2000 images from the Nordland datasets and calculate the average processing time. As expected, most CNN-based methods take generally more time to generate features with the exception of CALC and CoHOG. NetVLAD and Region-VLAD consume more time processing the descriptors than other approaches but gain on the matching performance. CALC and CoHOG sacrifice performance precision to obtain light-weight fea-

Table 5 Runtime comparing different methods on the Nordland Dataset.

VPR System	Feature Encoding (ms)	Feature Matching (ms)
Conv3	136	0.34
NetVLAD	980	0.038
CALC	39	0.31
HybridNet	158	0.36
CoHOG	31	0.29
Region-VLAD	700	0.61
FDNet	21.2	0.35
SFDNet	21.3	0.35

tures. Our proposed SFDNet does not change the CNN network structure for feature extraction compared with FDNet, thus the time consumed remains consistent.

6 Conclusion and future works

For visual place recognition with extremely dynamic conditions, achieving state-of-the-art performance using structure information is highly desirable but a challenging problem. This paper took a step in this direction and proposed a Structure-aware Feature Disentanglement Network (SFDNet) based on knowledge transfer and adversarial learning. The designed architecture derives novel CNN-based features that incorporate structure information extracted from the Canny edge detector through PKT. The proposed SFDNet achieves further improvements compared with the original FDNet, and achieved state-of-the-art AUC-PR curves and EP values using severe condition-variant datasets.

Future work will involve investigating how to alleviate the effect of changing viewpoints as well as exploring other types of appearance changes like dynamic objects.

Acknowledgements Supported by National Natural Science Foundation of China (No. 61973066, 61471110, 61733003) , National Key R&D Program of China(No.2017YFC0805000-/5005, 2017YFB1301103), Fundamental Research Funds for the Central Universities(N172608005, N182608004), Advanced Technology Project(No.41412050202), Natural Science Foundation of Liaoning (No.20180520040) and the Distinguished Creative Talent Program of Shenyang(RC170490)

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL http://tensorflow.org/. Software available from tensorflow.org
2. Anoosheh, A., Sattler, T., Timofte, R., Pollefeys, M., Van Gool, L.: Night-to-day image translation for retrieval-based localization. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 5958–5964. IEEE (2019)
3. Arandjelovic, R., Zisserman, A.: All about vlad. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
4. Arandjelović, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: Netvlad: Cnn architecture for weakly supervised place recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(6), 1437–1451 (2018)
5. Arroyo, R., Alcantarilla, P.F., Bergasa, L.M., Romera, E.: Are you able to perform a life-long visual topological localization? Autonomous Robots **42**(3), 665–685 (2018)
6. Bai, D., Wang, C., Zhang, B., Yi, X., Yang, X.: Sequence searching with cnn features for robust and fast visual place recognition. Computers & Graphics **70**, 270 – 280 (2018)
7. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6), 679–698 (1986)
8. Chen, Z., Jacobson, A., Sünderhauf, N., Upcroft, B., Liu, L., Shen, C., Reid, I., Milford, M.: Deep learning features at scale for visual place recognition. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 3223–3230. IEEE (2017)
9. Chen, Z., Liu, L., Sa, I., Ge, Z., Chli, M.: Learning context flexible attention model for long-term visual place recognition. IEEE Robotics and Automation Letters **3**(4), 4015–4022 (2018)
10. Cummins, M., Newman, P.: Fab-map: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research **27**(6), 647–665 (2008)
11. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893 vol. 1 (2005)
12. Eom, C., Ham, B.: Learning disentangled representation for robust person re-identification. In: Advances in Neural Information Processing Systems 32, pp. 5297–5308 (2019)
13. Facil, J.M., Olid, D., Montesano, L., Civera, J.: Condition-invariant multi-view place recognition. arXiv preprint arXiv:1902.09516 (2019)
14. Ferrarini, B., Waheed, M., Waheed, S., Ehsan, S., Milford, M.J., McDonald-Maier, K.D.: Exploring performance bounds of visual place recognition using extended precision. IEEE Robotics and Automation Letters **5**(2), 1688–1695 (2020)
15. Fukuda, T., Suzuki, M., Kurata, G., Thomas, S., Cui, J., Ramabhadran, B.: Efficient knowledge distillation from an ensemble of teachers. pp. 3697–3701 (2017). DOI 10.21437/Interspeech.2017-614
16. Gálvez-López, D., Tardos, J.D.: Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics **28**(5), 1188–1197 (2012)
17. Galvez-Lpez, D., Tardos, J.D.: Bags of binary words for fast place recognition in image sequences. IEEE Transactions on Robotics **28**(5), p.1188–1197 (2012)

18. Ge, Y., Li, Z., Zhao, H., Yin, G., Yi, S., Wang, X., Li, h.: Fd-gan: Pose-guided feature distilling gan for robust person re-identification. In: Advances in Neural Information Processing Systems 31, pp. 1222–1233 (2018)
19. Glover, A., Maddern, W., Milford, M., Wyeth, G.: FAB-MAP + RatSLAM: Appearance-based SLAM for Multiple Times of Day. In: ICRA. Anchorage, USA (2010)
20. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680 (2014)
21. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems, pp. 5767–5777 (2017)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). DOI 10.1109/CVPR.2016.90
23. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
24. Khaliq, A., Ehsan, S., Chen, Z., Milford, M., McDonald-Maier, K.: A holistic visual place recognition approach using lightweight cnns for significant viewpoint and appearance changes. IEEE Transactions on Robotics **36**(2), 561–569 (2020)
25. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
26. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., DENOYER, L., Ranzato, M.A.: Fader networks:manipulating images by sliding attributes. In: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (eds.) Advances in Neural Information Processing Systems 30, pp. 5967–5976 (2017)
27. Latif, Y., Garg, R., Milford, M., Reid, I.: Addressing challenging place recognition tasks using generative adversarial networks. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 2349–2355. IEEE (2018)
28. Laurens, V.D.M., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(2605), 2579–2605 (2008)
29. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Computer Vision – ECCV 2018, pp. 36–52 (2018)
30. Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J., Wang, X.: Exploring disentangled feature representation beyond face identification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2080–2089 (2018)
31. Lopez-Antequera, M., Gomez-Ojeda, R., Petkov, N., Gonzalez-Jimenez, J.: Appearance-invariant place recognition by discriminatively training a convolutional neural network. Pattern Recognition Letters **92**, 89 – 95 (2017)
32. Lowe, D.: Distinctive image features from scale-invariant key points. International Journal of Computer Vision **20**, 91–110 (2003)
33. Lowry, S., Milford, M.J.: Supervised and unsupervised linear learning techniques for visual place recognition in changing environments. IEEE Transactions on Robotics **32**(3), 600–613 (2016)
34. Lowry, S., Sünderhauf, N., Newman, P., Leonard, J.J., Cox, D., Corke, P., Milford, M.J.: Visual place recognition: A survey. IEEE Transactions on Robotics **32**(1), 1–19 (2015)
35. Lu, B., Chen, J.C., Chellappa, R.: Unsupervised domain-specific deblurring via disentangled representations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10225–10234 (2019)
36. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The oxford robotcar dataset. The International Journal of Robotics Research **36**(1), 3–15 (2017)
37. Mathieu, M.F., Zhao, J.J., Zhao, J., Ramesh, A., Sprechmann, P., LeCun, Y.: Disentangling factors of variation in deep representation using adversarial training. In: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (eds.) Advances in Neural Information Processing Systems 29, pp. 5040–5048 (2016)
38. Merrill, N., Huang, G.: Lightweight unsupervised deep loop closure. In: Proc. of Robotics: Science and Systems (RSS). Pittsburgh, PA (2018)
39. Merrill, N., Huang, G.: Lightweight unsupervised deep loop closure. In: Proc. of Robotics: Science and Systems (RSS). Pittsburgh, PA (2018)
40. Milford, M.J., Wyeth, G.F.: Seqslam: Visual route-based navigation for sunny summer days and stormy winter nights. In: 2012 IEEE International Conference on Robotics and Automation, pp. 1643–1649. IEEE (2012)
41. Naseer, T., Burgard, W., Stachniss, C.: Robust visual localization across seasons. IEEE Transactions on Robotics **34**(2), 289–302 (2018)
42. Naseer, T., Oliveira, G.L., Brox, T., Burgard, W.: Semantics-aware visual localization under challenging perceptual conditions. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 2614–2620. IEEE (2017)
43. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognition **29**(1), 51 – 59 (1996)
44. Passalis, N., Tefas, A.: Learning deep representations with probabilistic knowledge transfer. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
45. Passalis, N., Tzelepi, M., Tefas, A.: Probabilistic knowledge transfer for lightweight deep representation learning. IEEE Transactions on Neural Networks and Learning Systems pp. 1–10 (2020)
46. Porav, H., Maddern, W., Newman, P.: Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In: 2018 IEEE International Conference on Robotics and Automation (ICRA), pp. 1011–1018. IEEE (2018)
47. Qin, C., Zhang, Y., Liu, Y., Coleman, S., Kerr, D., Lv, G.: Appearance-invariant place recognition by adversarially learning disentangled representation. Robotics and Autonomous Systems p. 103561 (2020)
48. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550 (2014)
49. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: International Conference on Computer Vision (2012)
50. Rubner, Y., Tomasi, C., Guibas, L.J.: The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision **40**(2), 99–121 (2000)
51. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In:

2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
52. Sünderhauf, N., Neubert, P., Protzel, P.: Are we there yet? challenging seqslam on a 3000 km journey across all four seasons. In: Proc. of Workshop on Long-Term Autonomy, IEEE International Conference on Robotics and Automation (ICRA), p. 2013 (2013)
53. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4297–4304. IEEE (2015)
54. Sünderhauf, N., Shirazi, S., Jacobson, A., Dayoub, F., Pepperell, E., Upcroft, B., Milford, M.: Place recognition with convnet landmarks: Viewpoint-robust, condition-robust, training-free. Robotics: Science and Systems XI: pp. 1–10 (2015)
55. Sünderhauf, N., Shirazi, S., Dayoub, F., Upcroft, B., Milford, M.: On the performance of convnet features for place recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4297–4304 (2015)
56. Tang, L., Wang, Y., Luo, Q., Ding, X., Xiong, R.: Adversarial feature disentanglement for place recognition across changing appearance. In: 2020 International Conference on Robotics and Automation (ICRA), pp. 1301–1307. IEEE (2020)
57. Tang, Z., Wang, D., Zhang, Z.: Recurrent neural network training with dark knowledge transfer. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5900–5904 (2016)
58. Tzeng, E., Hoffman, J., Darrell, T., Saenko, K.: Simultaneous deep transfer across domains and tasks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4068–4076 (2015)
59. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7130–7138 (2017)
60. Yin, P., Xu, L., Li, X., Yin, C., Li, Y., Srivatsan, R.A., Li, L., Ji, J., He, Y.: A multi-domain feature learning method for visual place recognition. In: 2019 International Conference on Robotics and Automation (ICRA), pp. 319–324. IEEE (2019)
61. Zaffar, M., Ehsan, S., Milford, M., McDonald-Maier, K.: Cohog: A light-weight, compute-efficient, and training-free visual place recognition technique for changing environments. IEEE Robotics and Automation Letters **5**(2), 1835–1842 (2020)
62. Zaffar, M., Khaliq, A., Ehsan, S., Milford, M., McDonald-Maier, K.: Levelling the playing field: A comprehensive comparison of visual place recognition approaches under changing conditions. arXiv preprint arXiv:1903.09107 (2019)
63. Zeng, Z., Wang, Z., Wang, Z., Zheng, Y., Chuang, Y., Satoh, S.: Illumination-adaptive person re-identification. IEEE Transactions on Multimedia pp. 1–1 (2020)
64. Zhou, H., Liu, Y., Liu, Z., Luo, P., Wang, X.: Talking face generation by adversarially disentangled audio-visual representation. In: AAAI Conference on Artificial Intelligence (AAAI) (2019)